DOCUMENT RESUME

ED 346 138 TM 018 458

AUTHOR Isham, Steven P.; Allen, Nancy L.
TITLE New Information about Applications of Scale Anchoring.
PUB DATE 24 Apr 92
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; Criterion Referenced Tests; *Elementary School Students; Elementary Secondary Education; Item Response Theory; Multiple Choice Tests; National Norms; National Programs; *Rating Scales; *Scaling; Scores; *Secondary School Students; Student Evaluation; Testing Programs; *Test Items; Test Reliability; Trend Analysis
IDENTIFIERS *National Assessment of Educational Progress; Open Ended Questions; *Scale Anchoring

ABSTRACT

As a result of the dual roles of the National Assessment of Educational Progress (NAEP) to measure trends in academic achievement over time and to measure what students know and can do, a scale anchoring procedure was developed. Although the NAEP provides norm-referenced information about student proficiency, the scale anchoring procedure gives criterion-referenced information about the NAEP scales. Two studies compared results of scale anchoring the same items using data for three different grades and compared results of scale anchoring based on subscale proficiencies as opposed to composite proficiencies. In all, 275 items in 5 subscales were given to 19,251 students in grades 4, 8, and 12. Scale anchoring within each grade identifies slightly different sets of items as opposed to those identified by anchoring with all grade levels. However, only two items anchored at different levels for different grades. These items had different item characteristic curves for the different grades and were scaled separately for the different grades when the items were calibrated. Scale anchoring using subscales, as opposed to composite scales, also identifies slightly different sets of items. In both cases, items that anchor under one set of conditions but not another come very close to anchoring under the second set of conditions. There are nine tables of study data and an eight-item list of references. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED346138

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
NANCY L. ALLEN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

New Information about Applications of Scale Anchoring

Steven P. Isham and Nancy L. Allen

Educational Testing Service, Princeton, New Jersey

Paper presented at the annual meeting of the American Educational Research Association,

San Francisco, April 24, 1992

TM018458

BEST COPY AVAILABLE

Abstract

As a result of the dual roles of the National Assessment of Educational Progress (NAEP) to measure trends in academic achievement over time and to measure what students know and can do, a scale anchoring procedure was developed. Although NAEP provides norm-referenced information about student proficiency, the scale anchoring procedure gives criterion-referenced information about the NAEP scales. This paper describes two studies of the consistency of the scale anchoring procedures. Scale anchoring within each grade identifies slightly different sets of items as opposed to those identified by anchoring with all grade levels. However, only two items anchored at different levels for different grades. These items had different item characteristic curves for the different grades and were scaled separately for the different grades when the items were calibrated. Scale anchoring using subscales, as opposed to composite scales, also identifies slightly different sets of items. In both cases, items that anchor under one set of conditions but not another come very close to anchoring under the second set of conditions.

New Information about Applications of Scale Anchoring

From its inception, NAEP has had two complementary goals: to measure trends in academic achievement over time and to indicate what students in the U.S. know and can do. Beginning with the 1984 assessment, NAEP has used the comparison of scaled proficiency scores to measure trends across assessments, and scale anchoring procedures to give some indication of what students close to or at specific proficiency scale levels know and can do. Scaled proficiency scores have provided a means to look at norm-referenced results for the nation, while anchoring the proficiency scales has provided some criterion-referenced information about those results.

Currently the decision has been made that results from scale anchoring the traditional NAEP scale points (eg., 200, 250, 300, 350) will not be reported for the 1992 assessment. External achievement levels will be identified, however. These external achievement levels may or may not be anchored. Nevertheless, because the procedures can be applied to almost any set of test data, a comprehensive look at the validity evidence available for the methods in the NAEP setting is appropriate.

The scale anchoring procedures involve two components. The first component uses statistical techniques to identify exemplar items that have certain characteristics of interest. The second component uses a consensus process to interpret the results of the statistical component and to describe these interpretations in layman's terms. The results in this paper are based on different specifications for the statistical component of the anchoring procedures. This component of the procedures uses the item responses of students who are estimated to have scores at or near specified points on a scale to inditify items that might

indicate what students at the specified scale points know and can do. Each item is examined to see if it discriminates between successive specified scale points, called anchor points or anchor levels. The specified points on a scale are carefully selected so that they are far enough apart to identify different items and close enough to provide information about the scale in the actual range of scores. In the consensus component of the anchoring process, items that discriminate between adjacent anchor points are reviewed to see if their characteristics and the responses they require can be generalized to describe the level of proficiency at the anchor point.

The scale anchoring process is described in detail by Beaton and Allen (1992). The statistical component of the direct scale anchoring method is described here, however, because it was used in the analyses described in this paper. For each examinee, the pieces of information necessary to complete the process are right-wrong item scores and a scale score. In NAEP, "plausible values", generated so that secondary users of NAEP data can complete analyses of NAEP data (Mislevy, Johnson, & Muraki, 1992), are used in place of scale scores. Student weights are also used, because of NAEP's stratified sampling plan. Examinees whose scale scores are in a range around the specified scale points, the anchor levels, are grouped together. Examinees whose scale scores are not in the range around any of the anchor points do not contribute information to the scale anchoring process. The number of examinees in each group and the proportion of examinees in each group correctly answering each item are then calculated. For the lowest anchor point, the results for the examinees in the lowest group are examined. If (1) the proportion of examinees answering an item correctly exceeds a specified level and (2) the sample size of the group in the range

around the lowest anchor point is large enough, the item is said to anchor at the lowest anchor level. For successive anchor points, the proportion of examinees in the range around that point correctly answering an item is compared to the proportion for the range around the next lower anchor point. If the proportion for the range around the specified anchor point is large, the proportion for the range around the next lower anchor point is small, and the difference is large, then the item is said to anchor at the specified anchor level. Sample sizes for the groups of examinees in the ranges around the specified anchor point and the next lower anchor point are also checked to be sure that there are sufficient numbers of examinees in both groups. Several specifications must be made in the anchoring process, including:

- the examinees (data set) that will support the anchoring process,
- the scale that will be anchored,
- the set of items that will be anchored,
- the scale scores that will serve as anchor points,
- the range around the anchor points used to group examinees,
- the minimum sample size in the specified and next lower anchor point groups,
- the minimum proportion correct based on examinees in the specified anchor level group necessary for the item to anchor at a level,
- the maximum proportion correct based on examinees in the

next lower anchor level group necessary for the item to anchor at a level, and

- the minimum difference between the proportions correct based on examinees in the specified anchor level group and on examinees in the next lower anchor level group necessary for the item to anchor at the specified anchor level.

In previous studies, several aspects of the anchoring process have been examined. For instance, Mullis (1991) examined the consensus process by replicating that part of the scaling process for the 1990 mathematics assessment with two panels of mathematics curriculum experts. She found that the two panels agreed that their descriptions of the anchor points were "very similar and that, with some final review and editing, either set would have appropriately described the anchor item information" (Mullis, 1991, p. 6). Beaton and Allen (1992) replicated the statistical process for two random halves of the 1990 mathematics assessment. They found 83% exact agreement in the classification of the items. The remaining items anchored for one half-sample, but barely missed anchoring at the same level for the other half-sample. Beaton and Allen (1992) also evaluated the statistical component in a comparison of the direct method described above to a method in which the item characteristic curve for each item was smoothed. There was a 75% exact agreement between the two methods. As for the replication for half-samples of the data, the other items anchored using one method, but barely missed anchoring at the same level for the other method. These differences make little practical difference in the scale anchoring process, because items that almost anchor are included in the consensus process along with

items that do anchor and those that do not anchor. Other comparisons of anchoring results for different statistical specifications need to take place in order to know how to interpret scale anchoring results, and to present the results of the statistical component clearly and fairly to the panels participating in the consensus component of the process.

The purpose of this paper is two-fold: 1) to compare results of scale anchoring the same items using data for three different grades, and 2) to compare results of scale anchoring based on subscale proficiencies, as opposed to composite proficiencies. Forsyth (1991) and Koretz (1989), among others, take issue with the meaning of the scales implied by the results of the scale anchoring process. Until we know more about the statistical component of the scale anchoring process, discussion of the results of the consensus component is difficult and confusing. Even without a formal consensus process, these analyses provide information about the meaning of the NAEP scales by identifying items that anchor under one set of specifications, but not another. The interpretation of scale anchor points for NAEP and other tests should be influenced by whether scale anchoring results can be replicated for different grades and subscale proficiencies.

Methods

The NAEP 1990 mathematics main BIB (balanced incomplete block design) assessment provides an suitable data source for studying both the effect of grades and subscales on the anchoring process. The assessment consisted of 275 items in five subscales: (1) numbers and operations; (2) measurement; (3) geometry; (4) data analysis, statistics and probability; and (5) algebra and functions. These items were given to 19,251 students across three grade levels: four, eight and twelve. Of the 275 items, 99 contributed to the numbers

and operations subscale, 45 to the measurement subscale, 46 to the geometry subscale, 29 to the data analysis, statistics and probability subscale and 56 to the algebra and functions subscale. The data analysis, statistics and probability subscale was not designed for grade four. Therefore, there were no proficiency estimates for that scale at grade four.

At grade four, 77 multiple-choice and 28 open-ended items were presented. Of these 105 items, 45 were presented to at least one other grade. At grade eight, 101 multiple-choice and 35 open-ended were presented with 87 presented to at least one other grade. At grade twelve, 108 multiple-choice and 35 open-ended were presented with 64 items presented to at least one other grade. Since there was only one item common to grade four and grade twelve only, it was not included in the comparisons of items across grades. Objectives for the NAEP 1990 mathematics assessment are presented in *Mathematics Objectives, 1990 Assessment* (NAEP, 1988). Results are presented by Mullis, Dossey, Owen and Phillips (1991).

Despite the many choices for specifications that must be made in the anchoring process, the specifications for both the study of grade-only results and the study of subscale results differ only in two: the examinees that support the anchoring process and the scale that is anchored. For the first study, the composite proficiency is anchored for the students in each grade separately, as well as for all students together. For the second, each of the five subscale proficiencies, as well as the composite proficiency, were anchored using students from all three grades combined. The composite proficiency is a weighted average

of the five subscale proficiencies. Table 1 contains the composite proficiency weights used to combine the subscale proficiencies into a composite proficiency. These weights reflect the objectives for the NAEP 1990 mathematics assessment.

The remaining specifications are identical for both studies. Scale points 200, 250, 300 and 350 serve as the anchor points. A range of 25 points is used in grouping students' scores, with the anchor points serving as the midpoint of the range. The minimum sample size for the specified and next lower anchor level group is ten. At least 65% of the examinees in an anchor level group must respond correctly to an item, while fewer than 50% in the next lower anchor level group can respond correctly. The difference between the proportion correct at the two levels must be at least 30%. To be identified as anchoring at the 200 anchor level, an item must have at least 65% of the students grouped around the 200 score point responding correctly to it and the number of students in that group must be 10 or larger. To be identified as anchoring at the 250 anchor point, an item must have at least 65% of the students grouped around the 250 level responding correctly to it, while fewer than 50% of the students grouped around the 200 level can have responded correctly to the item. The difference between these two percentages must be at least 30%, and the number of students in each groups must be 10 or larger.

Included in the data are sample weights as well as scored item responses. These sampling weights are used in calculating the proportions correct. These estimates of the population values are unbiased. However, weights are not used in computing sample sizes. Because this paper is concerned with comparing specifications for the statistical component of the anchoring process, the consensus component is not replicated.

Results

Grade Level Comparisons

Table 2 contains the number of items anchoring at each anchor level for the within-grade samples and the across-grades sample. Applying the anchoring procedures to student from all grades that were assessed in the 1990 mathematics assessment identified 28 items that anchored at anchor level 200, 32 items that anchored at level 250, 52 items that anchored at level 300 and 33 items that anchored at level 350. Anchoring within each grade identifies a smaller set of items anchoring at each anchor level. This is not unexpected as the number of items administered to any particular grade is less than the total number of items presented in the assessment. For grade four, 24 items anchor at level 200, 22 items anchor at level 250, 12 items anchor at level 300 and no items anchor at level 350. For grade eight, items anchor at each of the levels with 10, 16, 25, and 12 items anchoring at levels 200, 250, 300 and 350, respectively. For grade twelve, no items anchor at levels 200 or 250, 36 items anchor at level 300 and 29 items anchor at 350. Anchor levels in which no items are identified are dropped before the consensus process. Therefore, given the current proficiency scale, the consensus component would arrive at a description of achievement at three levels for grade four and two levels for grade twelve.

Although anchoring within one grade does not identify a set of items that is identical to those identified for other grades, there are many items which are classified in the same way. The percent of exact agreement in the classification of items administered to more than one grade is presented in Table 3. For this analysis, items were classified as anchoring

at the 200, 250, 300 or 350 level, or as not anchoring at all. For each grade, the largest percent of exact agreement occurs when anchoring items that are presented to at least one other grade with students from all grades, as opposed to anchoring with students from one other grade. The percentages range from 70% exact agreement to 78% exact agreement. The lowest percent of exact agreement, 52%, occurs in two situations: anchoring grade four items that are also presented at grade twelve using twelfth graders' item responses and anchoring grade eight items that are also presented at grade twelve using twelfth graders' item responses. Similar to the findings of Beaton and Allen (1992), items that are not classified identically were very close to being classified in the same way. The items that are not identically classified were examined to determine if there are patterns to the reasons for the different classifications. There are no patterns. Items are classified in different ways due to small changes in the percent correct for the specified anchor level, small changes in the percent correct for the previous anchor level, small changes in the difference between the percent correct, or small changes in group sample sizes.

Table 4 contains the number of items presented in grades four and eight only that are classified in different ways using the grade four and grade eight data. Of these 24 items, 15 are identically classified either as an exemplar item for an anchor level or as not anchoring at all, and eight items anchor at an anchor level for one grade while not anchoring at any level at the other grade. One item anchors at level 250 for grade four and level 300 for grade eight. This item was treated as two separate items when IRT parameters were estimated, because the item response functions for grade four and grade eight were different (Yamamoto & Jenkins, 1992). A comparison of the classification for items

presented in grades eight and twelve only is presented in Table 5. Of the 42 items, 22 are identically classified and 20 anchor at a level for one grade while not anchoring at any level for the other grade. Again, items anchoring for one grade but not the other very narrowly missed anchoring at the same level for the other grade.

There is one item presented to all three grades that is similar to the item that anchors at different levels in grade four and eight. It anchors at anchor level 300 for grade four, at level 250 for grade eight, and does not anchor at grade twelve. However, it very nearly anchors at level 250 for grade twelve. Due to differing item response functions, grade four IRT parameters were estimated separately from grade eight and twelve parameters for this item. Out of the 21 items common to all three grades, there are six items that are identically classified for all three grades and 14 items that are classified in the same way for two grades, but not the third. None of the 14 items anchor at a different level for the remaining grade. Either the item anchors at the same level for two grades and does not anchor at the third or vice versa.

Subscale Comparisons

Table 6 contains the number of items anchoring at each anchor level when subscale proficiencies are used to anchor the 275 items presented in the 1990 mathematics assessment. Varying the scale being anchored identifies a set of items anchoring at every anchor level, unlike the case when grade levels are varied. The subscale proficiency that is weighted the most in the composite proficiency is numbers and operations, so it is not unexpected that anchoring with this subscale identifies items anchoring at each level in a way most similar to anchoring with the 1990 mathematics composite scale. There are 59

fewer items available for analysis when anchoring with the data analysis, statistics and probability subscale. The reason is that this proficiency was not estimated for grade four and there are 59 items administered only to fourth graders. So it is not unexpected that the fewest number of items anchoring at each level is associated with the data analysis, statistics and probability subscale.

As shown in Table 6, anchoring with each subscale does not identify identical sets of items that anchor at each anchor level. However, as in the case when anchoring within each grade, many items are classified in the same way. The percent of exact agreement of classification of items within a subscale using that subscale, as opposed to the composite scale or another subscale, is presented in Table 7. The percent of exact agreement in the classification of items when anchoring with a subscale as opposed to the composite scale ranges from 71% to 88%. For three subscales, numbers and operations, measurement and algebra and functions, the largest percent of exact agreement is with the composite scale. For the other two subscales, the percent of exact agreement is the same for both the composite scale and the numbers and operations subscale. Once again, items that are not classified in the same manner are close to being classified in the same way.

One might expect the percentage of exact agreement to follow the correlation of subscale proficiencies and the composite proficiency closely. The correlations are presented in Table 8. If the correlations in Table 8 and the percent agreement in Table 7 are ordered by magnitude, the order is exactly the same for two subscales and relatively close to being the same for the other three subscales. The orders of magnitude of the percent agreement for the numbers and operations and the algebra and functions subscales are identical to the

order of the correlations. For the measurement subscale, the percent agreement with the geometry subscale is two percent larger than the percent agreement with the algebra and functions subscale, whereas the correlations are the same. For the geometry subscale, the highest percent agreement is with the composite scale and the numbers and operations subscale followed by the algebra and functions subscale, whereas the correlation is highest for the composite scale followed by the numbers and operations and algebra and functions subscales. The order of the percent agreement for the data analysis, statistics and probability subscale is not as close to the order of the correlations as the other subscales. However, there are only 29 data analysis, statistics and probability items.

Table 9 contains a summary of the percent of exact agreement between anchoring with subscales as opposed to anchoring with the composite scale. In the first row, 32% of the 99 numbers and operations items are classified as anchoring at a certain level or as not anchoring for the numbers and operations subscale, for each of the other subscales and for the composite; 22% of the items are identically classified using the numbers and operation subscale and using the composite, but are not classified in the same way using any of the other subscales; 7% of the items were classified one way using the numbers and operations subscale and different ways using each of the other subscales and the composite; and 39% of the items show none of these patterns in the agreement of their classification across scales. The highest percent of items are classified in the same way for all subscales, or items that show no pattern to the classification across subscales. The former percentages range from 29% to 50%, while the latter ranges from 32% to 53%. Far fewer items are classified in the same way for the composite scale and the subscale of the item alone, 3% to 22%.

The percent of items for which the composite scale, and no other subscale, agrees in classification with the subscale of that item is lowest for all subscales, except for the data analysis, statistics and probability subscale.

Conclusion

Scale anchoring within grades, as opposed to across all three grades, and anchoring using subscales, as opposed to using the composite scale, did identify somewhat different sets of items at each anchor level. However, most items that are not classified in the same way barely miss being classified in the same way. Only two items anchored at different levels for different grades, and for both of these items, IRT item parameters were estimated separately for the different grades. In the two studies described, varying the students that support the anchoring process or the scales that are anchored identifies similar sets of items that anchor and almost anchor at a particular level. Because panelists in the consensus component of the anchoring procedure examine both items that anchor at a level and items that almost anchor at a level, it is reasonable to expect the description of achievement at a level to be similar under the different specifications described here.

References

Beaton, A. E. & Allen, N. L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics,* **, ***-***.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practices, 10*(3), 3-9, 16.

Koretz, D. (1989). *NAEP's scales: How useful are they?* Presentation at a meeting of the Education Commission of the States, Boulder, CO.

Mislevy, R. J., Johnson, E. G. & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics.* **, ***-***.

Mullis, I. V. S. (1991). *NAEP's 1990 mathematics scale: Anchor points and achievement levels.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Mullis, I. V. S., Dossey, J. A., Owen, E. H. & Phillips, G. W. (1991). *The state of mathematics achievement, NAEP's 1990 assessment of the nation and the trial assessment of the states.* Princeton, NJ: Educational Testing Service; Washington DC: National Center for Educational Statistics.

National Assessment of Educational Progress. (1988). *Mathematics objectives, 1990 assessment.* (No. 21-M-10) Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Yamamoto, K & Jenkins, F. (1992). Data analysis for the mathematics assessment. In E. G. Johnson, & N. L. Allen, *The NAEP 1990 technical report.* (No. 21-TR-20) Washington DC: National Center for Educational Statistics.

Table 1

Composite Proficiency Weights

Subscale	Grade 4	Grade 8	Grade 12
Numbers and Operations	50	30	25
Measurement	22	15	15
Geometry	17	20	20
Data Analysis, Statistics, and Probability	0	15	15
Algebra and Functions	11	20	25

Table 2

Number of Items Anchoring at Anchor Levels

	Number of Items	Anchor Level			
		200	250	300	350
All Grades	275	28	32	52	33
Grade 4	105	24	22	12	0
Grade 8	136	10	16	25	12
Grade 12	143	0	0	36	29

Table 3

Percent of Exact Agreement for Items Given to Multiple Grades

	Students Used in Anchoring Process		
	All Grades	Grade 12	Grade 8
Grade 4	78 (45)*	52 (21)**	60 (45)
Grade 8	74 (87)	52 (63)	
Grade 12	70 (63)		

*The number of items given at each grade level that are also given at least one other grade level is in parentheses.

**The number of items given at each grade level that are also given at the other grade level is in parentheses.

Table 4

Classification of the 24 Items Common to Grades Four and Eight

Grade 8 Anchor Level	Grade 4 Anchor Level			
	Does Not Anchor	200	250	300
Does Not Anchor	8		3	2
200	1	2		
250	1		3	
300	1		1	2

Table 5

Classification of the 42 Items Common to Grades Eight and Twelve

Grade 12 Anchor Level	Grade 8 Anchor Level				
	Does Not Anchor	200	250	300	350
Does Not Anchor	7	2	1	6	4
200					
250					
300	2			10	
350	5				5

Table 6

Number of Items Anchoring at Anchor Level

Subscale Proficiency Used in Anchoring All 275 Items	Anchor Level			
	200	250	300	350
Numbers and Operations	28	22	50	29
Measurement	34	7	28	15
Geometry	38	11	28	19
Data Analysis, Statistics and Probability*	21	5	19	14
Algebra and Functions	35	10	35	28

*Grade 4 students do not receive a data analysis, statistics and probability proficiency, therefore the 59 items that were presented only at grade 4 could not have anchored using this subscale proficiency in the scale anchoring process.

Table 7

Percent of Exact Agreement across Composite and Subscale Proficiencies

Scale of Item	Number of Items	Proficiency Used in Anchoring Process					
		Composite	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics and Probability	Algebra and Functions
Numbers and Operations	99	87		58	60	44	66
Measurement	45	71	69		60	53	58
Geometry	46	76	76	67		63	72
Data Analysis, Statistics and Probability	29	79	79	66	72		69
Algebra and Functions	56	88	77	57	59	54	

Table 8

Correlation of Composite and Subscale Proficiencies

	Composite	Numbers and Operations	Measurement	Geometry	Data Analysis, Statistics and Probability
Numbers and Operations	.96				
Measurement	.90	.82			
Geometry	.92	.83	.80		
Data Analysis, Statistics and Probability	.87	.76	.73	.73	
Algebra and Functions	.92	.85	.80	.83	.75

Table 9

Summary of Percent Agreement between Composite and Subscale Proficiencies

Scale of Items	Percent Agreement across All Subscales	Percent Agreement Only with Composite	Percent Agreement Only with Subscale Itself	No Pattern to the Percent Agreement
Numbers and Operations	32	22	7	39
Measurement	29	11	7	53
Geometry	50	9	9	32
Data Analysis, Statistics and Probability	45	3	10	42
Algebra and Functions	38	13	0	49